Ecologic Genomics of DNA: Upstream Bending in Prokaryotic Promoters

Alexander Bolshoy[1,2] and Eviatar Nevo[1]

[1]Institute of Evolution, University of Haifa, Haifa, 31905 Israel

After our analysis of the distribution of predicted intrinsic curvature along all available complete prokaryotic genomes, the genomes were divided into two groups. Curvature distribution in all prokaryotes of the first group indicated a substantial fraction of promoters characterized by intrinsic DNA curvature located within or upstream of the promoter region. We did not find this peculiar DNA curvature distribution in prokaryotes in the second group. Remarkably, all bacteria of the first group were mesophilic, whereas many prokaryotes of the second group were hyperthermophilic. We hypothesize that DNA curvature plays a biologic role in gene regulation in mesophilic as opposed to hyperthermophilic prokaryotes, i.e., DNA curvature presumably has a functional adaptive significance determined by temperature selection.

The determination of complete genome sequences led to evolutionary analysis at the comprehensive level of genomes. Computer analysis of complete prokaryotic genomes has resulted in characterization of families of orthologs across a wide phylogenetic range (Bork et al. 1998; Huynen and Bork 1998; Koonin et al. 1998), focusing primarily on gene evolution. Some recent research has studied the evolution of transcription regulation (Aravind and Koonin 1999; Gelfand et al., 2000). Our objective was to scrutinize, compare, and contrast gene regulation in Archaea and Bacteria. One such pattern of gene regulation is the presence of curved DNA upstream of a promoter, which has been described as "a common theme in prokaryotic gene expression" (Perez-Martin et al. 1994). The widely accepted hypothesis explaining the possible functional role of curvature in gene expression is that curved DNA assists in the formation of a large loop around RNA polymerase. Such a loop enhances the affinity of the complex to DNA and brings together components of the transcriptional complex that are otherwise more distant in the DNA sequence (Matthews 1992; Rippe et al. 1995). Curved DNA upstream to the promoter (upstream curved sequence, or UCS) has been shown to play a functionally regulatory role in *Escherichia coli* (Plaskon and Wartell 1987; Bracco et al. 1989; Lavigne et al. 1992; Carmona and Magasanik 1996; Dethiollaz et al. 1996).

In many of these and other investigations, presence of the curved DNA was established experimentally by a gel-electrophoretic anomaly technique. In many publications, existing computational models were shown to predict magnitude of DNA curvature with high reliability (Boffelli et al. 1992; Shpigelman et al. 1993; Goodsell and Dickerson 1994). In a previous study (Gabrielian et al. 1999), we applied the three most popular prediction models (De Santis et al. 1990; Bolshoy et al. 1991; Goodsell and Dickerson 1994) to compare distribution of average curvature values in different sets of sequences. One of the most important results of our study was that, qualitatively, all models demonstrated identical results. All three models indicated that UCS were found in *E. coli* substantially more frequently than it could be expected either from random distribution of DNA curvature along the genome or purely from A + T composition of noncoding DNA. We found that *E. coli* promoters as a set are significantly more curved than sets of coding sequences and randomized sequences (Gabrielian et al. 1999). *Escherichia coli* promoters also appeared to be more curved than randomly chosen fragments of *E. coli* noncoding sequences. In turn, noncoding sequences of *E. coli* were predicted to be more curved than coding and shuffled noncoding sequences. Interestingly, the robustness of the results was supported by the fact that in none of the three models was this effect found in the regions upstream to the human promoters.

In that study (Gabrielian et al. 1999), we took the opportunity to analyze well-developed databases of *E. coli* and human promoters. Unfortunately, locations of promoters are rarely established experimentally for other model organisms. However, in many cases, we were able to estimate a distance from a selected site to the nearest start of translation. This estimate might roughly indicate the relation of a method of selection to a promoter region. For example, we may select the most curved DNA fragments and study their distribution relatively to 5′ ends of predicted coding sequences (CDS). We used this approach in a previous work (Gabrielian and Bolshoy 1999), where we showed that features of distribution of putative UCS in *Bacillus subtilis*,

[2]**Corresponding author.**
E-MAIL bolshoy@esti.haifa.ac.il; FAX 972 4 8246554.

10:1185–1193 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

Haemophilus influenzae, and *Mycoplasma genitalium* are similar to those of *E. coli* DNA curvature distribution. Is this common genomic theme universal to all prokaryotic genomes?

To answer this question, we used statistical analysis. The fully annotated genomes provided essential information. We examined all complete prokaryotic genomes available through the Entrez browser provided at that time by the National Center for Biotechnology Information, six of which were euryarchaeal species and 15 bacteria. The consistent results of previous applications of different DNA curvature models (Gabrielian and Bolshoy 1999; Gabrielian et al. 1999) allowed us to select and apply only one such model for the purposes of the current study. The DNA curvature referenced to the *i*th base pair was measured as a value reciprocal to the radius of the arc fitting the predicted DNA path, with the center at the *i*th base pair (Shpigelman et al. 1993). The length of such an arc corresponds to the length of a DNA fragment, with the number of bases equal to the window size. We paid special attention to the curved regions predicted by use of the window size of 150 bases because putative DNA loops involved in transcription initiation have relatively large sizes. Thus, larger intrinsically bent fragments are the best candidates for such loops.

To study specificity of genomic curvature in intergenic regions, we applied a number of approaches. The first was to compare curvature distribution in contrasted sets of fragments, as in previous studies (Gabrielian and Bolshoy 1999; Gabrielian et al. 1999). In a previous study (Gabrielian and Bolshoy 1999), we showed that noncoding sequences of *Bacillus subtilis*, *Haemophilus influenzae,* and *Mycoplasma genitalium* are more curved than their corresponding control sequences. In the present study, we applied an approach analogous to all available complete prokaryotic genomes with a few window parameters. Our expectation was that results would be qualitatively independent of the window size, similar to what was shown for *E. coli* by Gabrielian et al. (1999).

In the present study, we describe the results of detailed comparative analysis of the distribution of predicted intrinsic curvature. Together, these results point to a presumably adaptive environmental division: mesophilic versus hyperthermophilic species.

RESULTS

Comparison of Average Curvature Values of Coding and Noncoding Genomic Sequences

It was repeatedly shown that, in prokaryotic organisms, intergenic regions are on average more curved than coding regions (VanWye et al. 1991; Gabrielian et al. 1997; Jauregui et al. 1998; Gabrielian and Bolshoy 1999). Table 1 shows the average curvature values of noncoding, coding, and shuffled noncoding regions. The curvature values were measured in nucleosome units (n.u.), as explained by Shpigelman et al. (1993). The largest values are shown in boldface type. Note that results are very robust relative to the window size. The noncoding sequences were found almost everywhere to be more curved than corresponding CDS; the exception was *Treponema pallidum*. We calculated the significance probability values (*p*) by the TTEST procedure (see Methods). This procedure showed that, in *Aquifex aeolicus*, *Methanococcus jannaschii*, and *Thermotoga maritima*, the hypothesis about random distribution of curvature along the genome could not be rejected. Because our expectation was that DNA curvature should be concentrated mainly in noncoding regions, this simple analysis casts doubt on a biologic role for DNA curvature in the life cycles of these four species.

Average Curvature Values of Noncoding Genomic Sequences and Random Sequences with Identical Base Composition

This comparison was performed to test the hypothesis that intergenic base composition per se may explain the differences in average curvature. Data in Table 1 appear to reject this hypothesis. Indeed, Table 1 provides evidence that, in almost every complete prokaryotic genome, noncoding (intergenic) sequences are more curved than their shuffled counterparts. However, we found exceptions: *Methanobacterium thermoautotrophicum* (0.074 vs. 0.078) and *Aeropyrum pernix* (0.062 vs. 0.064). These average curvature values correspond to the window size of 150 bp. The measurements seem to show that higher values of upstream gene curvature are avoided in these two prokaryotic genomes. Curvature distributions of noncoding sequences of *A. aeolicus*, *T. pallidum*, and *T. maritima* were practically indistinguishable from those of their shuffled counterparts. It is interesting to compare average curvature values of genomes with similar AT composition. For example, *Pyrococcus horikoshii* has an AT composition of about 58% and *Chlamydia pneumonia* of about 59%. The average curvature values of noncoding DNA are fairly distinct, despite the fact that the values of AT composition are very close. The corresponding values in Table 1 are 0.137 versus 0.151 (window size equal to 50 bp), 0.097 versus 0.109 (window size of 100 bp), and 0.080 versus 0.089 (window size of 150 bp).

Curved DNA in the Neighborhood of the CDS

The second approach was to study curvature distribution around the 5′ ends of the CDS. Our expectation was that UCS would be located upstream to starts of translation in regions of putative promoters. For every genome we calculated the curvature distributions in

Table 1. Average DNA Curvature Genomic Values

			Average curvature (n.u.)[a] Noncoding vs. coding vs. shuffled noncoding								
Genome	Size (bp)	% AT	window = 50 bp			window = 100 bp			window = 150 bp		
Escherichia coli	4639221	49	**0.141**	0.119	0.129	**0.104**	0.084	0.092	**0.087**	0.068	0.075
Mycobacterium tuberculosis	4411529	34	**0.099**	0.089	0.096	**0.071**	0.063	0.068	**0.059**	0.052	0.056
Bacillus sp.	4214814	56	**0.153**	0.132	0.135	**0.111**	0.093	0.094	**0.092**	0.076	0.077
Synechocystis PCC6803	3573470	52	**0.154**	0.137	0.129	**0.112**	0.098	0.091	**0.093**	0.080	0.074
Haemophilus influenzae	1830138	62	**0.170**	0.145	0.143	**0.124**	0.102	0.101	**0.104**	0.083	0.082
Helicobacter pylori	1667867	61	**0.178**	0.160	0.147	**0.133**	0.115	0.104	**0.113**	0.094	0.084
Chlamydia pneumonia	1230230	59	**0.151**	0.135	0.140	**0.109**	0.096	0.099	**0.089**	0.079	0.081
Chlamydia trachomatis	1042519	59	**0.153**	0.135	0.134	**0.109**	0.096	0.096	**0.090**	0.079	0.077
Mycoplasma pneumoniae	816394	60	**0.156**	0.140	0.137	**0.109**	0.097	0.095	**0.089**	0.079	0.076
Mycoplasma genitalium	580074	68	**0.161**	0.151	0.143	**0.113**	0.107	0.100	**0.088**	0.088	0.081
Borrelia burgdorferi	910724	71	**0.166**	0.160	0.149	**0.123**	0.114	0.103	**0.104**	0.094	0.083
Rickettsia prowazekii	1111523	71	**0.158**	0.154	0.151	**0.114**	0.109	0.107	**0.095**	0.089	0.089
Treponema pallidum	1138011	47	**0.118**	**0.117**	0.107	0.080	**0.083**	0.080	0.065	**0.068**	0.065
Hyperthermophiles											
Aquifex aeolicus	1551335	57	**0.140**	**0.136**	0.133	**0.100**	**0.097**	0.093	**0.081**	**0.080**	0.076
Thermotoga maritima	1860725	54	**0.136**	0.126	0.131	**0.095**	0.090	**0.093**	**0.076**	0.073	**0.076**
Archaea											
Aeropyrum pernix	1669695	44	**0.109**	0.100	**0.109**	**0.077**	0.071	**0.077**	0.062	0.058	**0.064**
Pyrococcus horikoshii	1738505	58	**0.137**	0.130	0.133	**0.097**	0.092	0.095	**0.080**	0.076	0.077
Pyrococcus abyssi	1765118	55	**0.138**	0.123	0.131	**0.098**	0.087	0.092	**0.082**	0.071	0.074
Archaeoglobus fulgidus	2178400	51	**0.143**	0.128	0.131	**0.101**	0.092	0.091	**0.083**	0.075	0.073
Methanobacterium thermoautotrophicum	1751377	50	0.130	0.111	**0.135**	0.091	0.079	**0.096**	0.074	0.065	**0.078**
Methanococcus jannaschii	1664970	69	**0.159**	0.152	0.145	**0.113**	0.108	0.102	**0.092**	0.088	0.082

[a]The largest values are shown in boldface type.

the neighborhood of the 5′ ends (±500 nucleotides) and averaged these distributions over all such fragments. Figure 1 shows all curves on the same scale but are shifted along the *y* axis for illustration purposes. The curvature values are presented in nucleosome units, as in Table 1. Major ticks on the *y* axis correspond to the curvature of 0.02 nucleosome unit. We divided the genomes into three groups: big mesophilic genomes (Fig. 1A), small mesophilic genomes (Fig. 1B), and all hyperthermophilic genomes (Fig. 1C).

Six mesophilic genomes (Fig. 1A) clearly express nonrandom distribution, with curvature peaks between 100 and 200 bases upstream to the starts of the CDS. *Mycobacterium tuberculosis* distribution has less variance than that of other mesophilic genomes; however, it evidently has larger curvature upstream rather than downstream to the CDS, with the maximum in the region of −125 to −275 relative to the starts of the CDS.

Among seven smaller mesophilic genomes (Fig. 1B), five plots show upstream asymmetry similar to that of larger mesophilic bacteria, albeit without convincing statistical significance. *Treponema pallidum* and *Rickettsia prowazekii* did not demonstrate any preference of curvature distribution.

Five of eight hyperthermophilic genomes (Fig. 1C) have no obvious preference in curvature distribution. Three others are *A. pernix*, *M. thermoautotrophicum*, and *P. abyssi*. We discussed average curvatures of noncoding DNA in the genomes of *M. thermoautotrophicum* and *A. pernix* above. Reshuffling of noncoding DNA of these genomes can even increase average curvature. Thus, it is reasonable to suggest that larger curvature of upstream regions in these two hyperthermophiles is mainly a consequence of the noncoding AT composition. The upstream curvature in *P. abyssi* is probably of a different nature and plays a functional role. In this respect, *P. abyssi* is a special case: its curvature distribution curve is different from that of other hyperthermophiles and is more similar to that of mesophilic bacteria, such as *B. subtilis*.

In addition to the average curvature distribution around starts of translation, we investigated the distribution of locations of maximum curvature values in the same fragments (data not shown). This complementary approach produced consistent results.

Length Distribution of Intergenic Regions

After our analysis of the curvature distribution, we clustered the genomes into two groups. In principle, such clustering may be the result of an artifact: differences in length distribution of intergenic regions in different genomes can produce artifacts. Table 2 presents an extraction from the genomic annotations,

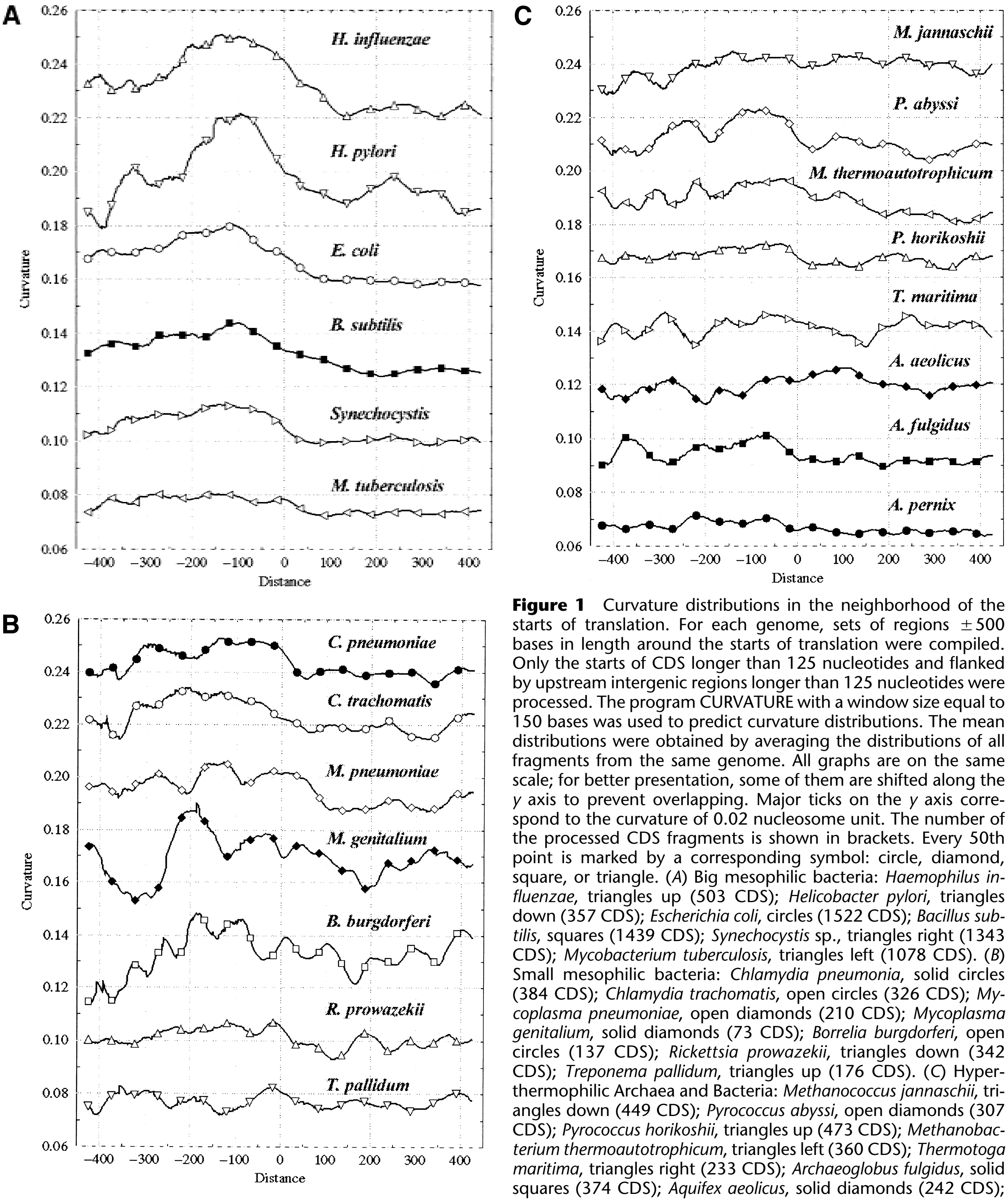

Figure 1 Curvature distributions in the neighborhood of the starts of translation. For each genome, sets of regions ±500 bases in length around the starts of translation were compiled. Only the starts of CDS longer than 125 nucleotides and flanked by upstream intergenic regions longer than 125 nucleotides were processed. The program CURVATURE with a window size equal to 150 bases was used to predict curvature distributions. The mean distributions were obtained by averaging the distributions of all fragments from the same genome. All graphs are on the same scale; for better presentation, some of them are shifted along the *y* axis to prevent overlapping. Major ticks on the *y* axis correspond to the curvature of 0.02 nucleosome unit. The number of the processed CDS fragments is shown in brackets. Every 50th point is marked by a corresponding symbol: circle, diamond, square, or triangle. (*A*) Big mesophilic bacteria: *Haemophilus influenzae*, triangles up (503 CDS); *Helicobacter pylori*, triangles down (357 CDS); *Escherichia coli*, circles (1522 CDS); *Bacillus subtilis*, squares (1439 CDS); *Synechocystis* sp., triangles right (1343 CDS); *Mycobacterium tuberculosis*, triangles left (1078 CDS). (*B*) Small mesophilic bacteria: *Chlamydia pneumonia*, solid circles (384 CDS); *Chlamydia trachomatis*, open circles (326 CDS); *Mycoplasma pneumoniae*, open diamonds (210 CDS); *Mycoplasma genitalium*, solid diamonds (73 CDS); *Borrelia burgdorferi*, open circles (137 CDS); *Rickettsia prowazekii*, triangles down (342 CDS); *Treponema pallidum*, triangles up (176 CDS). (*C*) Hyperthermophilic Archaea and Bacteria: *Methanococcus jannaschii*, triangles down (449 CDS); *Pyrococcus abyssi*, open diamonds (307 CDS); *Pyrococcus horikoshii*, triangles up (473 CDS); *Methanobacterium thermoautotrophicum*, triangles left (360 CDS); *Thermotoga maritima*, triangles right (233 CDS); *Archaeoglobus fulgidus*, solid squares (374 CDS); *Aquifex aeolicus*, solid diamonds (242 CDS); *Aeropyrum pernix*, solid circles (749 CDS).

being a brief description of sizes of intergenic regions. For every prokaryotic genome, we show both absolute and relative numbers of occurrences. Size distributions are similar for almost all genomes, although there are a few exceptions. The most extraordinary genome is that of *Rickettsia prowazekii*. The *R. prowazekii* genome contains the highest proportion of noncoding DNA (24%) detected thus far in a microbial genome (Andersson et al. 1998). Not only *R. prowazekii* but also *A. pernix* intergenic regions are longer, on average, than those in other genomes. The opposite is true regarding the *A. fulgidus*, *T. pallidum*, and *P. abyssi* genomes: the average

Table 2. Sizes of Intergenic Regions[a]

Organism	≤25	≤125	≤475	≤1000	>1000	All
Aeropyrum pernix	0	426	560	201	107	1324
	0.0	0.322	0.425	**0.153**	0.081	
Archaeoglobus fulgidus	46	661	318	45	51	1121
	0.041	0.590	0.284	0.041	0.045	
Aquifex aeolicus	32	289	178	41	11	551
	0.058	0.524	0.323	0.075	0.020	
Borrelia burgdorferi	33	231	113	9	5	391
	0.084	0.591	0.288	0.023	0.013	
Bacillus sp.	112	1366	1262	133	51	2924
	0.038	0.467	0.431	0.046	0.017	
Chlamydia pneumonia	29	260	326	27	15	657
	0.044	0.395	0.496	0.042	0.023	
Chlamydia trachomatis	32	219	266	16	9	542
	0.059	0.404	0.491	0.031	0.017	
Escherichia coli	86	1222	1281	169	45	2803
	0.031	0.436	0.456	0.060	0.016	
Haemophilus influenzae	25	620	424	47	61	1177
	0.021	0.527	0.361	0.041	0.052	
Helicobacter pylori	45	383	316	36	49	829
	0.054	0.462	0.382	0.042	0.059	
Mycoplasma genitalium	9	85	53	12	15	174
	0.052	0.488	0.303	0.071	0.086	
Mycoplasma pneumoniae	15	110	147	56	25	353
	0.042	0.312	0.415	**0.157**	0.071	
Methanococcus jannaschii	46	530	399	46	55	1076
	0.043	0492	0.371	0.045	0.051	
Methanobacterium thermoautotrophicum	55	590	401	26	9	1081
	0.051	0.546	0.370	0.024	0.008	
Mycobacterium tuberculosis	92	1203	902	125	77	2399
	0.038	0.501	0.377	0.051	0.032	
Pyrococcus abyssi	31	499	236	57	29	852
	0.036	0.586	0.277	0.066	0.034	
Pyrococcus horikoshii	24	432	364	95	97	1012
	0.024	0.427	0.360	0.095	0.096	
Rickettsia prowazekii	14	202	231	84	167	698
	0.020	0.290	0.331	**0.121**	**0.239**	
Synechocystis PCC6803	62	1204	1205	127	13	2611
	0.024	0.461	0.462	0.049	0.005	
Thermotoga maritima	34	339	155	24	51	603
	0.056	0.562	0.257	0.041	0.085	
Treponema pallidum	28	359	177	8	19	591
	0.047	0.608	0.300	0.015	0.032	

[a]The largest values are shown in boldface type.

sizes of intergenic regions are smaller than those of the other genomes. We mentioned that curvature distribution profiles of *T. pallidum* and *P. abyssi* genomes are dissimilar. It is most unlikely that deviation to smaller intergenic regions has the opposite effect on *T. pallidum* and *P. abyssi* genomes. Our general conclusion from the analysis shown in Table 2 is that size distribution of intergenic regions does not correlate with curvature distribution and does not produce meaningful artifacts. Nevertheless, we used another approach that was independent of the sizes of intergenic regions to investigate curvature distribution.

Location of the Most Curved Regions in the Genome

Whereas the first two methods distinguished curvature of coding from noncoding regions, the third one studies distribution of the most curved pieces along an entire genome. We used a three-step procedure that was applied to all available complete prokaryotic genomes. The first step was to produce curvature maps of an entire genome, with window sizes of 63, 150, 250, and 400 bp. The second step was to find a threshold such that regions with curvature over the threshold together cover close to a predefined part of genome length (we used values of 1.5% and 5% of the genome length) and to compile sets of all such pieces for every window separately. The third step was to construct histograms of distances from centers of the selected curved regions to the nearest annotated CDS. The histograms obtained for *E. coli* and *H. influenza* with the threshold equal to 1.5% are shown in Fig. 2. The main result presented in Fig. 2 is that, for all window sizes, the most curved regions are preferentially placed about 100–200 bases upstream to the nearest CDS. Other "big" mesophilic genomes, namely *B. subtilis*, *M. tuberculosis*, *H. pylori*, and *Synechocystis*, have distribution profiles of the most curved pieces very similar to those of *E. coli* and *H. influenza*. Note that *M. tuberculosis* is placed among the "curved" genomes despite a very low A + T composition and low average curvature. The results obtained by this CDS-independent approach generally confirmed the results achieved by different methods.

DISCUSSION

Evolutionary Conservation of DNA Loops in Bacteria

In many bacterial promoters, certain upstream sequences were able to stimulate higher transcription in vivo by factors of tens or even hundreds (e.g., Lamond and Travers 1983; Perez-Martin and de Lorenzo 1997; Ross et al. 1993, 1998). These upstream sequences frequently appeared to exhibit inherent DNA curvature.

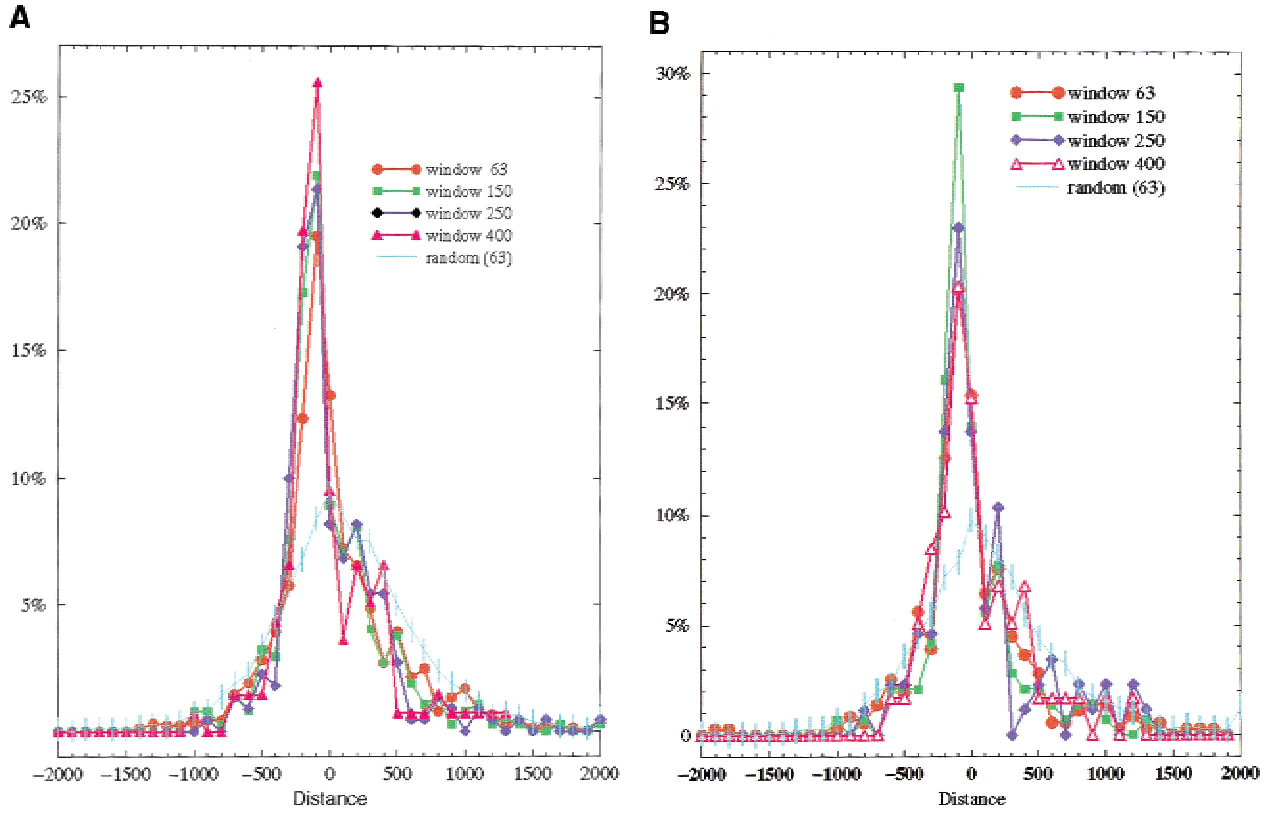

Figure 2 Distances from the centers of the most curved regions to the nearest starts of translation. All regions with curvature over the threshold together cover ~1.5% of genome length. The curvature maps with window sizes of 63, 150, 250, and 400 bp were calculated. The histograms of distances from the centers of the selected regions to the nearest annotated CDS are shown. In addition, distances from randomly generated points to the CDS were calculated. The window size of 63 bp is presented by a line with circles, 150 bp by squares, 250 bp by diamonds, and 400 bp by triangles. (*A*) *Escherichia coli K-12*. The number of the most curved fragments selected is 886 for window size 63 bp, 370 for size 150 nucleotides, 220 for 250 bp, and 137 for 400 bp. (*B*) *Haemophilus influenzae*. The number of the most curved fragments selected is 357 for window size 63 bp, 143 for 150 nucleotides, 87 for 250 bp, and 59 for 400 bp.

However, the precise mechanism of action of these bacterial enhancers is still unknown. The role of curved DNA has been questioned. For example, when studying the rrnB P1 promoter, Aiyar et al. (1998) suggested an explanation for the different effects of upstream A-tracts on transcription versus those of the intrinsic curvature caused by phased A-tracts. The conservative patterns of genomic curvature distribution across different mesophilic bacterial genomes presented in the present study provide a new, convincing indication that curved DNA is evolutionarily preserved. There are two major mechanisms for curved DNA to assist in a loop formation: directly or as a binding site for a structure-recognizing bending protein. The most curved fragments of a genome, measured by a large window of 250 and 400 bp, are the best candidates for direct assistance in a loop formation. Figure 2 shows that more than half (!) of these loop candidates in *E. coli* and *H. influenza* are located in presumed promoter areas. We observed a similar effect in other big mesophilic genomes. Our conclusion is that preferable location of long curved fragments upstream to transcription initiation starts is evolutionarily preserved and related to a mesophilic environment.

Curved DNA and Temperature Influence

Correlation between classification of genomes according to their curvature distribution and thermophility is not coincidental. Temperature and other environmental influences on the intrinsic DNA curvature, expressed as an electrophoretic anomaly, have been studied (e.g. Diekmann 1987; Ussery et al. 1999). Chan et al. (1993), by using a variety of physical methods, also detected a temperature-dependent, “premelting” event that eliminates DNA curvature, and they suggested that this event corresponds to the specific DNA structure of curved DNA. It was universally found that the effect of DNA curvature disappears with rising temperature. This relationship between temperature and DNA curvature may suggest functional significance of DNA curvature in hyperthermophilic Archaea, not under physiologic conditions but at lower temperatures only. We found that the hypothesis of Lopez-Garcia (1999) corresponds very well with our findings. Lopez-

Garcia speculated that distinctive DNA topology in hyperthermophilic Archaea appeared as a result of evolution and that it participates in gene regulation in response to environmental changes. In that context (Lopez-Garcia 1999), the term “DNA topology” is almost synonymous with “DNA supercoiling”. In the context of our work, DNA curvature is a simplified characteristic of the overall DNA structure. Nevertheless, the conclusions of Lopez-Garcia are very close to ours. This agreement is not surprising, because DNA topology and DNA spatial trajectory are directly related. Interestingly, the observation that transcription in *Pyrococcus* hyperthermophiles seems to be regulated differently from that of other hyperthermophiles (Bell et al. 1998; Soares et al. 1998) agrees well with our observations.

Archaeal Transcription and Role of DNA Curvature in Gene Regulation

It was indicated that transcription in Archaea is more homologous to that in Eukarya than to that in Bacteria (Bell and Jackson 1998; Bell et al. 1998). Such homologues as between eukaryotic and archaeal TATA-binding proteins and between basal transcription factors TFIIB and TFB were mentioned. Curved DNA was implied among common DNA structural features exhibited by eukaryotic ribosomal gene promoters (Marilley and Pasero 1996). In a previous study (Gabrielian et al. 1999), we tried to answer the question of whether intrinsic DNA curvature is a necessary component of human promoter architecture. We found that, in eukaryotes, the frequency of occurrence of curved fragments in human promoter sequences hardly exceeds that for coding regions. We concluded that the regulation of eukaryal transcription rarely involves DNA curvature. In the present study, we suggest that the same is true with respect to archaeal transcription. As in eukaryotes, we assume that upstream curved sequences in hyperthermophiles are rarely involved in initiation of transcription. Presumably, UCS may be used by hyperthermophilic prokaryotes in response to cold shock (Bell et al. 1998).

The possibility exists that a well-established similarity between eukaryal and archaeal transcriptional apparatus causes similarity in the regulation of eukaryal and archaeal transcription. In particular, this may be an origin of the aforementioned similarity with respect to a role of DNA curvature. However, the hyperthermophilic bacteria *Aquifex aeolicus* and *Thermotoga maritima* do not express patterns of preferences in DNA curvature distribution as do Archaea. The transcriptional apparatus of *A. aeolicus* is similar to that of *E. coli* and lacks components specific to Archaea (Deckert et al. 1998). The same probably holds for *T. maritima* transcription. Environmental stress in hyperthermophiles may well cause evolutionary deficiency of UCS and transcription factors binding to curved DNA.

With regard to analyzing the seven graphs shown in Fig. 1B, we mentioned that five plots show upstream asymmetry similar to that of larger mesophilic bacteria, whereas *Treponema pallidum* and *Rickettsia prowazekii* do not show any preference in curvature distribution. We speculate that this DNA curvature deficiency in *T. pallidum* and *R. prowazekii* is of a nature different than that of hyperthermophiles. *Treponema pallidum* has indistinguishably low values of DNA curvature everywhere (Table 1): only *M. tuberculosis* and *A. pernix* express lower values. Moreover, the curvature average values of *T. pallidum* coding sequences are higher than those of noncoding sequences. This makes the *T. pallidum* genome an exception among practically all completely sequenced genomes. Seemingly, the syphilis spirochete does not use DNA curvature as a gene regulation factor. *Rickettsia prowazekii* may tell a different story. It has such an unusual genome structure and such a large fraction of noncoding DNA (Table 2; Andersson et al. 1998) that our statistical methods could not investigate this genome.

Concluding Remarks

All applied methods of curvature distribution analysis showed a substantial presence of more curved pieces 100–200 bases upstream to the start of CDS in such mesophilic genomes as *Escherichia coli*, *Mycobacterium tuberculosis*, *Bacillus* sp., *Synechocystis*, *Haemophilus influenzae*, and *Helicobacter pylori*. In contrast, both euryarchaeal and bacterial hyperthermophilic species did not demonstrate such a property. DNA curvature probably does not play any significant biologic role in the gene regulation of hyperthermophilic species. Our analysis does not determine unambiguously whether this lack of significant upstream curvature is typical for *Treponema pallidum* and *Rickettsia prowazekii* or whether it is an artifact produced by unusual genome structures. Further investigations should clarify this question. We predict that analysis of new complete prokaryotic genomes as a rule will show patterns of curvature distribution that are dependent on normal growth temperatures.

METHODS

Curvature Calculation

There are several models and methods of DNA curvature calculation. However, in previous publications (Gabrielian and Bolshoy 1999; Gabrielian et al. 1999), we demonstrated that different methods of curvature calculation were found to produce mostly similar overall tendencies of DNA curvature in all groups of sequences. In the present study, the prediction of DNA curvature was made by means of our CURVATURE program. This program calculates a three-dimensional path of a DNA molecule and estimates the curvature of the axis path

(Shpigelman et al. 1993) with dinucleotide wedge angles (Bolshoy et al. 1991).

Extraction of Coding and Noncoding Regions

Automatic procedures of extraction used annotations of complete prokaryotic genome sequences in GenBank, release 111. For every genome, one set of coding fragments and two sets of noncoding pieces were obtained by these procedures. The set of coding fragments consists of 250-bp-long pieces randomly chosen from every annotated CDS. The first set of noncoding pieces deals with all intergenic regions at least 100 bp in length, and the second set deals with those 250 bp and longer. The first set was processed by a window size of 21 bp and consisted of fixed-length regions of 250 bp each, immediately upstream. The second set included complete intergenic regions and was processed by a window size of 150 bp. Control "random" sets were obtained from corresponding noncoding sequences by reshuffling of sequences.

In our study of curvature distribution around the 5′ ends of the CDS, we processed only CDS longer than 125 nucleotides and flanked by upstream intergenic regions longer than 125 nucleotides. We aimed to take a neighborhood of ±500 bases in length. However, entire regions ±500 bases in length around the starts of translation were used exclusively in cases where all 500 nucleotides upstream were located in an intergenic region and all 500 bases downstream belonged to CDS. Otherwise, only relevant downstream coding or upstream noncoding pieces were used.

Dispersion Analysis

To test the hypothesis that noncoding sequences have mean curvature distribution values different from those of coding and shuffled sequences, we used the TTEST and MEANS procedures of the SAS software. The TTEST procedure computes *t* statistic based on the assumption that the variances of coding and noncoding groups are unequal. This is an approximate *t* statistic for testing the null hypothesis that the means of the two groups are equal. The probability value *p* (Table 1) is the probability of greater absolute value of *t* under the null hypothesis. The TTEST procedure provided the two-tailed significance probability, and the MEANS procedure was used to obtain paired-comparison *t* tests between noncoding and shuffled noncoding sequences. The same procedure was applied to test the hypothesis that the distribution of distances from the peak of curvature in the neighborhood of the start of translation to the start is normally symmetric.

ACKNOWLEDGMENTS

We thank A. Gabrielian, E.N. Trifonov, and V. Kirzhner for discussions and critical comments on the paper. Thanks are due to A. Beharav for his assistance in using SAS programs for the dispersion analysis. A.B. is a Guastaella Fellow. This work was funded in part by the Israel Discount Bank Chair of Evolutionary Biology and the Ancell-Teicher Research Foundation for Genetics and Molecular Evolution.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Aiyar, S.E., Gourse, R.L., and Ross, W. 1998. Upstream A-tracts increase bacterial promoter activity through interactions with the RNA polymerase alpha subunit. *Proc. Natl. Acad. Sci. U.S.A.* **95:** 14652–14657.

Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T., Alsmark, U.C., Podowski, R.M., Naslund, A.K., Eriksson, A.S., Winkler, H.H., and Kurland, C.G. 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature* **396:** 133–140.

Aravind, L. and Koonin, E.V. 1999. DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res.* **27:** 4658–4670.

Bell, S.D. and Jackson, S.P. 1998. Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. *Trends Microbiol.* **6:** 222–228.

Bell, S.D., Jaxel, C., Nadal, M., Kosa, P.F., and Jackson, S.P. 1998. Temperature, template topology, and factor requirements of archaeal transcription. *Proc. Natl. Acad. Sci. U.S.A.* **95:** 15218–15222.

Boffelli, D., De Santis, P., Palleschi, A., Risuleo, G., and Savino, M. 1992. A theoretical method to predict DNA permutation gel electrophoresis from the sequence. *FEBS Lett.* **300:** 175–178.

Bolshoy, A., McNamara, P., Harrington, R.E., and Trifonov, E.N. 1991. Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc. Natl. Acad. Sci. U.S.A.* **88:** 2312–2316.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. 1998. Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283:** 707–725.

Bracco, L., Kotlarz, D., Kolb, A., and Diekmann, S., and Buc, H. 1989. Synthetic curved DNA sequences can act as transcriptional activators in *Escherichia coli. EMBO J.* **8:** 4289–4296.

Carmona, M. and Magasanik, B. 1996. Activation of transcription at sigma 54-dependent promoters on linear templates requires intrinsic or induced bending of the DNA. *J. Mol. Biol.* **261:** 348–356.

Chan, S.S., Breslauer, K.J., Austin, R.H., and Hogan, M.E. 1993. Thermodynamics and premelting conformational changes of phased (dA)5 tracts. *Biochemistry* **32:** 11776–11784.

De Santis, P., Palleschi, A., Savino, M., and Scipioni, A. 1990. Validity of the nearest-neighbor approximation in the evaluation of the electrophoretic manifestations of DNA curvature. *Biochemistry* **29:** 9269–9273.

Deckert, G., Warren, P.V., Gaasterland, T., Young, W.G., Lenox, A.L., Graham, D.E., Overbeek, R., Snead, M.A., Keller, M., Aujay, M. 1998. The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. *Nature* **392:** 353–358.

Dethiollaz, S., Eichenberger, P., and Geiselmann, J. 1996. Influence of DNA geometry on transcriptional activation in *Escherichia coli. EMBO J.* **15:** 5449–5458.

Diekmann, S. 1987. Temperature and salt dependence of the gel migration anomaly of curved DNA fragments. *Nucleic Acids Res.* **15:** 247–265.

Gabrielian, A.E., and Bolshoy, A. 1999. Sequence complexity and DNA curvature. *Comput. Chem.* **23:** 263–274.

Gabrielian, A.E., Landsman, D., and Bolshoy, A. 1999. Curved DNA in promoter sequences. *In Silico Biol.* **1:** 0017; http://www.bioinfo.de/isb/1999/01/0017.

Gabrielian, A., Vlahovicek, K., and Pongor, S. 1997. Distribution of sequence-dependent curvature in genomic DNA sequences. *FEBS Lett.* **406:** 69–74.

Gelfand, M.S., Koonin, E.V., and Mironov, A.A. 2000. Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res.* **28:** 695–705.

Goodsell, D.S. and Dickerson, R.E. 1994. Bending and curvature calculations in B-DNA. *Nucleic Acids Res.* **22:** 5497–5503.

Huynen, M.A. and Bork, P. 1998. Measuring genome evolution. *Proc. Natl. Acad. Sci. U.S.A.* **95:** 5849–5856.

Jauregui, R., O'Reilly, F., Bolivar, F., and Merino, E. 1998. Relationship between codon usage and sequence-dependent curvature of genomes. *Microb. Comp. Genom.* **3:** 243–253.

Koonin, E.V., Tatusov, R.L., and Galperin, M.Y. 1998. Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* **8:** 355–363.

Lamond, A.I. and Travers, A.A. 1983. Requirement for an upstream element for optimal transcription of a bacterial tRNA gene. *Nature* **305:** 248–250.

Lavigne, M., Herbert, M., Kolb, A., and Buc, H. 1992. Upstream curved sequences influence the initiation of transcription at the *Escherichia coli* galactose operon. *J. Mol. Biol.* **224:** 293–306.

Lopez-Garcia, P. 1999. DNA supercoiling and temperature adaptation: a clue to early diversification of life? *J. Mol. Evol.* **49:** 439–452.

Marilley, M. and Pasero, P. 1996. Common DNA structural features exhibited by eukaryotic ribosomal gene promoters. *Nucleic Acids Res.* **24:** 2204–2211.

Matthews, K.S. 1992. DNA looping. *Microbiol. Rev.* **56:** 123–136.

Perez-Martin, J. and de Lorenzo, V. 1997. Clues and consequences of DNA bending in transcription. *Annu. Rev. Microbiol.* **51:** 593–628.

Perez-Martin, J., Rojo, F., and de Lorenzo, V. 1994. Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. *Microbiol. Rev.* **58:** 268–290.

Plaskon, R.R. and Wartell, R.M. 1987. Sequence distributions associated with DNA curvature are found upstream of strong *E. coli* promoters. *Nucleic Acids Res.* **15:** 785–796.

Rippe, K., von Hippel, P.H., and Langowski, J. 1995. Action at a distance: DNA-looping and initiation of transcription. *Trends Biochem. Sci.* **20:** 500–506.

Ross, W., Aiyar, S.E., Salomon, J., and Gourse, R.L. 1998. *Escherichia coli* promoters with UP elements of different strengths: modular structure of bacterial promoters. *J. Bacteriol.* **180:** 5375–5383.

Ross, W., Gosink, K.K., Salomon, J., Igarashi, K., Zou C., Ishihama, A., Severinov, K., and Gourse, R.L. 1993. A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science* **262:** 1407–1413.

Shpigelman, E.S., Trifonov, E.N., and Bolshoy, A. 1993. CURVATURE: software for the analysis of curved DNA. *Comput. Appl. Biosci.* **9:** 435–440.

Soares, D., Dahlke, I., Li, W.T., Sandman, K., Hethke, C., Thomm, M., and Reeve, J.N. 1998. Archaeal histone stability, DNA binding, and transcription inhibition above 90 degrees C. *Extremophiles* **2:** 75–81.

Ussery, D.W., Higgins, C.F., and Bolshoy, A. 1999. Environmental influences on DNA curvature. *J. Biomol. Struct. Dyn.* **16:** 811–823.

VanWye, J.D., Bronson, E.C., and Anderson, J.N. 1991. Species-specific patterns of DNA bending and sequence. *Nucleic Acids Res.* **19:** 5253–5261.

Received August 31, 1999; accepted in revised form June 2, 2000.